**Instagram Reach Analysis**

Project submitted for the partial fulfillment of the requirements for the course

**CSE 336L: Machine Learning Lab**

Offered by the

**Department Computer Science and Engineering School of Engineering and Sciences**

Submitted by

1)Karlapudi Charishma, Roll No: AP2111001026 8

2)Sravani Katragadda, Roll No: AP21110010293

3)Deekshitha Chirumamilla, Roll No: AP21110010304



SRM University–AP

**Neerukonda, Mangalagiri, Guntur Andhra Pradesh – 522 240 [May, 2024]**

**Contents:**

|  |  |  |
| --- | --- | --- |
| **SNO** | **TOPIC** | **PAGE NO** |
| **1.** | **Introduction** | **3** |
| **2.** | **Background** | **3** |
| **3.** | **Data set** | **4** |
| **4.** | **procedure** | **5** |
| **5.** | **ML models** | **8** |
| **6.** | **Results** | **10** |
| **7.** | **Conclusion** | **11** |

# Introduction:

# In the digital age, social media platforms have become essential channels for individuals and businesses to connect with their audiences. Instagram has become a crucial platform for individuals and businesses to connect and engage with their audiences. With over a billion active users worldwide, understanding the dynamics of post reach on Instagram is essential. The reach of a post is influenced by various factors, including content characteristics, audience demographics, and engagement patterns. Analyzing these factors presents a complex challenge that can benefit from the application of data science techniques and machine learning algorithms.

This project aims to develop a predictive model that estimates the potential reach of Instagram posts based on their attributes. The study will use machine learning and data analysis approaches to uncover the factors that affect reach. Two regression models, the Random Forest Regressor and the GradientBoost Regressor, will be used to make accurate predictions by capturing complex relationships within the data. Evaluation metrics such as Mean Squared Error (MSE) and R-squared will be used to assess the performance of the models.

This compares the Random Forest and GradientBoost Regressors in predicting post reach on Instagram. The insights gained from this analysis can help develop a predictive model and understand the underlying mechanisms of post-reach dynamics on social media platforms. The ultimate goal is to empower users and businesses to optimize their content strategies and enhance their impact on Instagram. Through meticulous analysis and innovative modeling techniques, the project aspires to unlock new avenues for harnessing the power of social media influence in the digital age.

## Background:

Social media has transformed communication and community-building for individuals and businesses. Instagram is a major player in this arena, providing a visually rich platform for expression and engagement. It is essential for content creators, influencers, marketers, and brands to utilize Instagram to reach their target audience and generate engagement. Post reach on Instagram is determined by various factors, including content type, timing, audience demographics, engagement metrics, and algorithmic factors.

The field of data science has seen significant advancements in machine learning algorithms and predictive modeling techniques. Researchers and practitioners are exploring social media dynamics to develop predictive models that can accurately forecast post reach and engagement. This project focuses specifically on Instagram post reach prediction, aiming to uncover the underlying factors driving post reach and provide actionable insights to inform content creators, marketers, and brands. By bridging the gap between data science methodologies and social media analytics, stakeholders can thrive in the ever-evolving landscape of social media influence.

## Data set:

This dataset serves as the foundation for our analysis and predictive modeling efforts aimed at understanding the factors influencing the reach of Instagram posts and developing a predictive model to estimate post reach based on its attributes. The features included in our dataset are as follows:

**1.** **Impressions:** This represents the total number of impressions generated by a post, indicating its reach or the number of times it was displayed on users' screens.

**2. From Home:** The reach of the post originating from users' home feed.

**3. From Hashtags:** The reach of the post attributed to users discovering it through hashtags.

**4. From Explore:** The reach of the post from users exploring content on the Explore page.

**5. From Other:** The reach of the post from sources other than the home feed, hashtags, or Explore page.

**6. Saves:** The number of times the post was saved by users.

**7. Comments:** The number of comments received on the post.

**8. Shares:** The number of times the post was shared by users.

**9. Likes:** The number of likes received on the post.

**10. Profile Visits:** The number of visits to the profile of the account posting the content.

**11. Follows:** The number of users who followed the account as a result of the post.

**12. Caption:** The text accompanying the post, providing context or additional information.

**13. Hashtags:** The hashtags used in the post to categorize and increase its discoverability.

The dataset for Instagram posts provides valuable insights into performance and engagement metrics. Through advanced machine learning techniques, a predictive model will be developed to estimate the potential reach of Instagram posts based on their attributes. The aim is to empower users and businesses with the knowledge and tools needed to enhance their impact and maximize their reach on the platform.

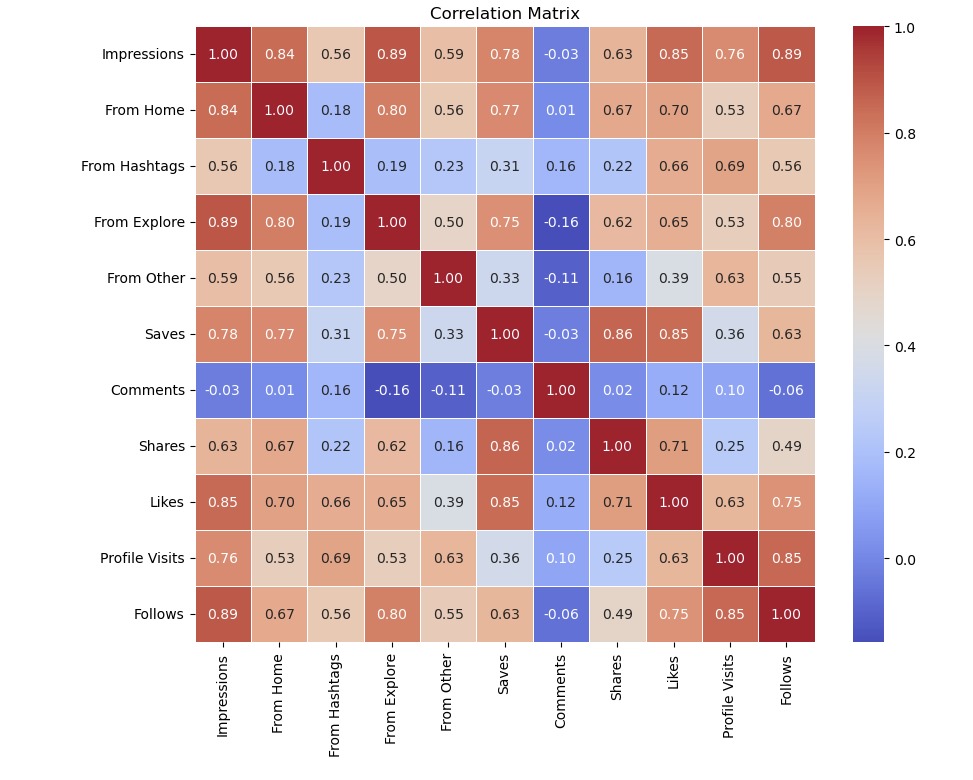
**Procedure:**

1.Data preprocessing:

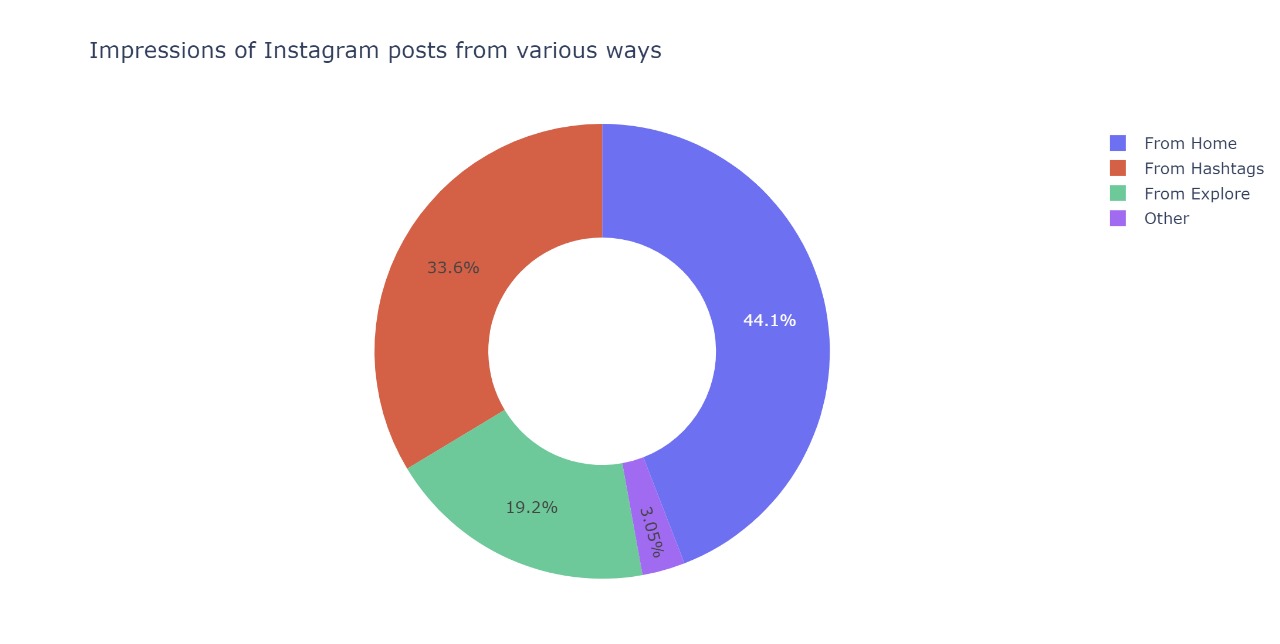
* **Handle missing values:** Impute missing values in numerical features with median values. Replace missing values in categorical features (e.g., captions, hashtags) with "unknown".
* **Text data processing:** Tokenize captions and hashtags. Remove punctuation and convert text to lowercase.
* **Encode categorical variables:** One-hot encode categorical features like post reach source (e.g., From Home, From Hashtags).
* **Train-test split:** Split the dataset into training and testing sets while maintaining data distribution.

2. Co-relation analysis:

We conducted a correlation analysis to explore the relationships between various numerical variables related to post engagement and reach on Instagram. The correlation matrix revealed the degree of association between different features, providing valuable insights into potential dependencies and patterns within the dataset. By examining this correlation matrix, we identified significant correlations between certain engagement metrics, such as likes, comments, shares, and profile visits, which can serve as potential predictors of post-reach. These insights are crucial for building our predictive model and understanding the key drivers of post-reach and engagement on the platform.

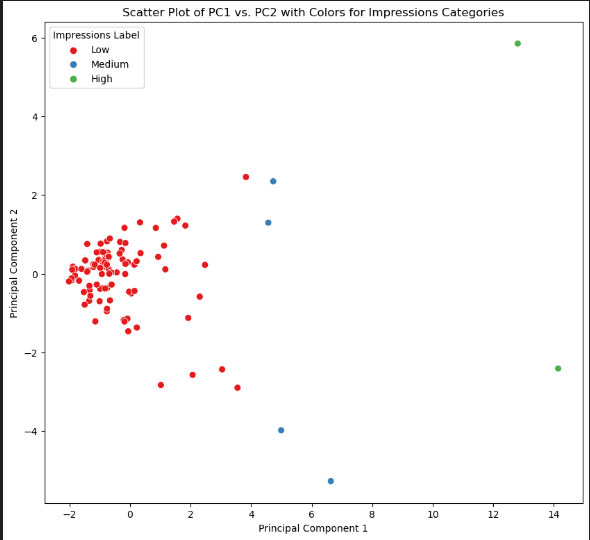


3.Data visualization:  
We present a bar plot that visually ranks the correlation coefficients of various numerical variables with Impressions on Instagram posts. The height of each bar indicates the strength and direction of the correlation, with positive coefficients suggesting an increase in Impressions as the variable value increases, and negative coefficients indicating an inverse relationship. This plot helps to identify the most strongly associated features with post reach and guides our feature selection process for building a predictive model.



4.Principal Component Analysis:

A Principal Component Analysis (PCA) was performed on engagement metrics of a social media platform to understand underlying patterns. The dataset was standardized to ensure comparability across variables, and the PCA revealed that four principal components captured over 90% of the total variance. A scatter plot was used to visualize data points in the reduced-dimensional space defined by the first two principal components, color-coded based on impressions levels (Low, Medium, or High). The analysis provides insights for enhancing user engagement and content performance on the platform.



## Machine Learning Models:

## 1.Gradient Boosting Regressor:

## A Gradient Boosting Regressor model was utilized to analyze the correlation between engagement metrics and impressions on social media. The model employs ensemble learning by fitting weak learners to the residuals of the previous model, minimizing prediction error. Features such as 'From Explore', 'Follows', 'Likes', and others were used to train the model, with the target variable being 'Impressions'. The Gradient Boosting Regressor captures complex interactions between features and predicts impression levels accurately. Mean squared error and R-squared were used to assess the model's performance. Insights gained from the analysis can aid in decision-making processes, optimizing engagement strategies and content creation efforts on social media.

## 

## 2.Random Forest Regressor:

## A Random Forest Regressor was utilized to analyze the link between engagement metrics and impressions on a social media platform. The model was trained on a dataset consisting of various features with the target variable being impressions. The Random Forest Regressor is effective in capturing complex dynamics of social media engagement and mitigates overfitting by averaging multiple trees' predictions, leading to reliable predictions of impression levels. Evaluation metrics such as MSE or R-squared were used to validate the model's predictive capabilities. The insights gained from the model can inform strategic decision-making processes, optimizing engagement strategies and content performance on the social media platform.

## 

## 3. Support vector Regression:

## Support Vector Regression (SVR) was employed as a predictive modeling technique to analyze the connection between engagement metrics and impressions on social media. SVR is a supervised learning algorithm that identifies the optimal hyperplane in a high-dimensional feature space to represent the relationship between input features and the target variable. The SVR model was trained on a dataset consisting of various features like 'From Explore', 'Follows', 'Likes', 'From Home', 'Saves', 'Profile Visits', 'Shares', 'From Other', and 'From Hashtags', with 'Impressions' as the target variable. SVR excels at capturing nonlinear relationships between features and the target variable, making it ideal for modeling complex interactions in social media engagement data. By employing a kernel function, SVR can identify nonlinear patterns and correlations by mapping input features into a higher-dimensional space. Evaluation metrics such as mean squared error (MSE) or R-squared were used to assess the model's performance and validate its predictive accuracy. Insights obtained from SVR can offer valuable guidance for optimizing engagement strategies and improving content performance on social media platforms.

## 

# Results:

# The comparison analysis assessed the performance of three regression models: Gradient Boosting Regressor, Random Forest Regressor, and Support Vector Regression. Evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) were used to measure predictive accuracy, precision, and goodness-of-fit. The results showed that Gradient Boosting Regressor outperformed the other models in all metrics, indicating superior predictive accuracy and precision. It also had the highest R-squared value, suggesting a better explanation of the variance in impressions levels. While Random Forest Regressor performed well, it fell slightly short of Gradient Boosting Regressor. Support Vector Regression had higher MSE and RMSE values, indicating lower predictive accuracy. In conclusion, using Gradient Boosting Regressor is recommended for predicting impressions on social media due to its superior performance. This emphasizes the importance of advanced machine learning techniques in optimizing engagement strategies and content performance.

# 

# Conclusion:

Our analysis delved into the complex relationship between engagement metrics and impressions on social media. By utilizing advanced machine learning techniques including Gradient Boosting Regressor, Random Forest Regressor, and Support Vector Regression, we gained valuable insights into the factors that impact impression levels. Among these models, the Gradient Boosting Regressor proved to be the most effective in accurately predicting impressions. This highlights the importance of engagement metrics like likes, shares, and profile visits in driving impressions on the platform. These insights can empower content creators and marketers to optimize their engagement strategies, expand their audience reach, and achieve their marketing objectives in the ever-evolving digital landscape